DOCUMENT RESUME

ED 211 061 IR 009 878

AUTHOR	Tatsuoka, Kikumi; Linn, Robert L.
TITLE	Indices for Detecting Unusual Item Response Patterns in Personnel Testing: Links Between Direct and Item-Response-Theory Approaches. Computerized Adaptive Testing and Measurement.
INSTITUTION	Illinois Univ., Urbana. Computer-Based Education Research Lab.
SPONS AGENCY	Office of Naval Research, Washington, D.C. Psychological Sciences Div.
REPORT NO	CERL-RR-81-5
PUB DATE	Aug 81
CONTRACT	N000-14-79-C-0752
NOTE	38p.; For related document, see IR 009 879.
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Arithmetic; *Computer Assisted Testing; *Latent Trait Theory; *Matrices; *Statistical Analysis; Testing Problems
IDENTIFIERS	*Response Patterns; *S P Curve Theory

ABSTRACT

Two distinct approaches, one based on item response theory and the other based on observed item responses and standard summary statistics, have been proposed to identify unusual response patterns. A link between these two approaches is provided by showing certain correspondences between Sato's S-P Curve Theory and item response theory. This link makes possible several extensions of Sato's caution index that take advantage of the results of item response theory. Several such indices are introduced and their use illustrated by application to a set of achievement test data. Two of the newly introduced extended indices were found to be very effective for purposes of identifying persons who consistently use an erroneous rule in attempting to solve signed number arithmetic problems. The potential importance of this result is briefly discussed, and 15 references are listed. (Author/LLS)

University of Illinois
Computer-based Education Research Laboratory
Urbana, Illinois

INDICES FOR DETECTING UNUSUAL ITEM RESPONSE PATTERNS IN PERSONNEL TESTING: LINKS BETWEEN DIRECT AND ITEM-RESPONSE-THEORY APPROACHES

by

Kikumi Tatsuoka

and

Robert L. Linn

Computerized Adaptive Testing and Measurement
Research Report 81-5
August 1981

Acknowledgment

The authors wish to acknowledge the kind cooperation extended to us by the people involved with this report. Bob Baillie programmed the lessons and data collection and analysis routines, along with his assistant, David Dennis. Mary Klein gave insight and meaning to many things as a teacher of the children whom we seek to help. Roy Lipschutz did the layouts and Louise Brodie did the typing.

Abstract

Two distinct approaches, one based on item-response theory and the other based on observed item responses and standard summary statistics, have been proposed to identify unusual response patterns. A link between these two approaches is provided by showing certain correspondences between Sato's S-P curve Theory and item response theory. This link makes possible several extensions of Sato's caution index that take advantage of the results of item response theory. Several such indices are introduced and their use illustrated by application to a set of achievement test data. Two of the newly introduced extended indices were found to be very effective for purposes of identifying persons who consistently use an erroneous rule in attempting to solve signed-number arithmetic problems. The potential importance of this result is briefly discussed.

Introduction

Several authors have recently shown an interest in using information from patterns of response to test items to extract information not contained in the total score. A variety of purposes have been envisioned for use of the additional information. Wright (1977), for example, refers to identification of "guessing, sleeping, fumbling, and plodding" (p. 110) from the plots of residual item scores based on the differences between item responses and the expected responses for an individual based on the Rasch model. Levine and Rubin (1979) discuss response patterns that are "so atypical... that his or her aptitude test score fails to be a completely appropriate measure" (p. 269). Sato (1975) proposed a "caution" index which is intended to identify students whose total scores on a test must be treated with caution. Tatsuoka and Tatsuoka (1980) and Harnisch and Linn (1981) have discussed the relationship of response patterns to instructional experiences and the possible use of item response pattern information to help diagnose the types of errors a student is making.

Indices of the degree to which an individual's pattern of responses is unusual are conveniently classified into two general types: those that use item response theory (IRT) to identify unusual patterns and those that rely only on observed item responses and standard summary statistics based on those responses (e.g. the number or proportion of people in a norm group answering an item correctly). The work of Wright (1979) and of Levine and Rubin (1979) are examples of approaches based on IRT while the work of Sato (1975), Tatsuoka and Tatsuoka (1980), and Harnisch and Linn (1981) are of the latter type.

The primary purpose of this paper is to develop a link between these two general approaches. More specifically, we will show a correspondence between Sato's (1975) S-P Curve theory and test response curves and "group response curves" developed from IRT. Also, Sato's Caution Index defined in the S-P curve theory is generalized into a continuous domain utilizing IRT. That is, S-P curve theory and the Caution Index are originally developed in a discrete domain of 0 - 1 scoring, but this study extends the theory to a more general case of probabilities.

Several different generalized versions of the caution index are presented. Results of applying these indices suggest that there are two categories. One set of indices functions in a manner similar to Sato's original index. The other set functions more like Tatsuoka and Tatsuoka's Individual Consistency Index in that it successfully distinguishes examinees who make consistent errors in responding to test items.

We first briefly review Sato's S-P Curve theory. Next, a group response curve (GRC) is developed for the one parameter logistic model. The GRC is based on the dualistic nature of the one parameter logistic model which depends on the choice of fixed and random parameters in the model. We then present an extended caution index with several special areas which are applicable to IRT. The cases of two and three parameter logistic models are briefly discussed with special attention given to problems with person and group response curves in these models. Finally, we discuss applications of the new caution indices for the detection of anomalous response patterns.

S-P Curve Theory

Sato's (1975) caution index is applicable to either an item or an individual examinee. In either form the index is conveniently obtained from an especially arranged table of binary item scores referred to as an "S-P Table". The S-P table, the associated S-P curves and various indices as the caution index are widely used in Japan for diagnosing student performance, detecting aberrant response patterns and for assessing the quality of a test or instructional sequence.

The S-P table is a data matrix in which the students (represented by rows) have been arranged in descending order of their total test scores from top to bottom and the items (represented by columns) have been arranged in ascending order of difficulty from left to right. A hypothetical S-P table is shown in Table 1. The solid stair-step line is called an S-curve which is short for Student curve. For each person, represented by a given row, a vertical line is drawn to the right of the nth cell from the left where n is the number of correct answers obtained by that person. The S-curve is then obtained by connecting the right edge of the nth cell of each row. The P-curve is drawn in an analogous fashion by counting down from the top the number of cells equal to the number of students who correctly answered the item corresponding to a given column. The P-curve for the data in Table 1 is shown by the dashed line.

Insert Table 1 about here

Let y_{ij} be the binary response for student (row) i to item (column) j of the S-P table. Row and column sums are denoted by $y_{i.}$ and $y_{.j}$ respectively. The total number of ones in the S-P table is denoted $y_{..}$

Table 1
A Hypothetical Score Matrix (y_{ij}) and
S- (solid line) and P- (dotted line) Curves

item j / subject i	1	2	3	4	5	6	7	8	9	10	$y_{i.}$	$P_{i.}$	$M^S_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10	1.0	10
2	1	1	1	1	1	1	1	1	1	0	9	0.9	9
3	1	1	1	1	1	0	1	1	0	1	8	0.8	8
4	1	0	1	1	1	1	0	1	0	0	6	0.6	6
5	1	1	1	1	0	1	0	0	1	0	6	0.6	6
6	1	1	1	0	1	0	1	0	1	0	6	0.6	6
7	1	1	1	1	0	0	1	0	0	0	5	0.5	5
8	1	1	1	0	1	1	0	0	0	0	5	0.5	5
9	1	0	0	1	0	1	0	1	1	0	5	0.5	5
10	1	1	0	1	0	0	1	0	0	1	5	0.5	5
11	0	1	1	1	1	0	0	0	0	0	4	0.4	4
12	1	0	0	0	1	1	0	0	0	1	4	0.4	4
13	1	1	0	0	0	1	0	0	0	0	3	0.3	3
14	1	0	1	0	0	0	0	0	0	0	2	0.2	2
15	0	1	0	0	0	0	0	0	0	0	1	0.1	1
$y_{.j}$	13	11	10	9	8	8	6	5	5	4		$y_{..}$ = 79	
$P_{.j}$	.87	.73	.67	.60	.53	.53	.40	.33	.33	.27		$P_{..}$ = .527	
$M^P_{.j}$	13	11	10	9	8	8	6	5	5	4			

and the proportion of correct responses by $P_{i.}$, $P_{.j}$ and $P_{..}$ for the row, column and entire table respectively. As can be seen in Table 1, the S-curve is the step function ogive of the mulative distribution function of total scores, $y_{i.}$, for the 15 students and the P-curve is the corresponding function of $y_{.j}$, the number of right answers for the 10 items.

Insert Table 2 about here

If the S-curve is held invariant and all the 0's to the left of the S-curve are changed to 1's and all the 1's to the right of the same curve to 0's the result is the S-P table shown in Table 2 is called a perfect S-curve. The entries in Table 2 are denoted M^s_{ij}. Similarly a perfect P-curve will be obtained and the entries in the new table are denoted by M^p_{ij}. As can be seen, $M^s_{i.} = y_{i.}$ for all i which corresponds to the fact that the S-curve is unchanged as the result of changing the cell entries from y_{ij} to M^S_{ij}. The values of the column sums for Tables 1 and 2, i.e., $y_{.j}$ and $M^P_{.j}$ are not in general equal, however.

Sato (1975) defined a Caution Index for subject i by taking the ratio of two covariances. The numerator of the ratio is the covariance of observed row vector i, (y_{ij}) j=1,...,n and the sum-of-column vector, $(y_{.j})$, j=1,2,...,n and the denominator is the covariance of the corresponding scores (assuming S-curve is perfect) (M^s_{ij}), j=1,...,n and the column-sum vector $(y_{.j})$, j=1,2,...,n. More specifically, the caution index C_i for the subject i is given by

$$C_i = 1 - \frac{\sum_{j=1}^{n}(y_{ij} - P_{i.})(y_{.j} - P_{..})}{\sum_{j=1}^{n}(M^s_{ij} - P_{i.})(y_{.j} - P_{..})}$$

Table 2
Perfect S-curve Obtained by Changing 1's to the Right of S-curve to 0 and 0's to the Left to 1.

item j / subject i	1	2	3	4	5	6	7	8	9	10	$y_{i.}$	$M^S_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	0	9	9
3	1	1	1	1	1	1	1	1	0	0	8	8
4	1	1	1	1	1	1	0	0	0	0	6	6
5	1	1	1	1	1	1	0	0	0	0	6	6
6	1	1	1	1	1	1	0	0	0	0	6	6
7	1	1	1	1	1	0	0	0	0	0	5	5
8	1	1	1	1	1	0	0	0	0	0	5	5
9	1	1	1	1	1	0	0	0	0	0	5	5
10	1	1	1	1	1	0	0	0	0	0	5	5
11	1	1	1	1	0	0	0	0	0	0	4	4
12	1	1	1	1	0	0	0	0	0	0	4	4
13	1	1	1	0	0	0	0	0	0	0	3	3
14	1	1	0	0	0	0	0	0	0	0	2	2
15	1	0	0	0	0	0	0	0	0	0	1	1
$y_{.j}$	13	11	10	9	8	8	6	5	5	4	79	79
$M^P_{.j}$	15	14	13	12	10	6	3	3	2	1		

Perfect P-Curve Obtained by Changing 1's Below the P-Curve to 0 and 0's Above to 1

item j / subject i	1	2	3	4	5	6	7	8	9	10	$y_{i.}$	$M^S_{i.}$
1	1	1	1	1	1	1	1	1	1	1	10	10
2	1	1	1	1	1	1	1	1	1	1	9	10
3	1	1	1	1	1	1	1	1	1	1	8	10
4	1	1	1	1	1	1	1	1	1	1	6	10
5	1	1	1	1	1	1	1	1	1	0	6	9
6	1	1	1	1	1	1	1	0	0	0	6	7
7	1	1	1	1	1	1	0	0	0	0	5	6
8	1	1	1	1	1	1	0	0	0	0	5	6
9	1	1	1	1	0	0	0	0	0	0	5	4
10	1	1	1	0	0	0	0	0	0	0	5	3
111	1	1	0	0	0	0	0	0	0	0	4	2
12	1	0	0	0	0	0	0	0	0	0	4	1
13	1	0	0	0	0	0	0	0	0	0	3	1
14	0	0	0	0	0	0	0	0	0	0	2	0
15	0	0	0	0	0	0	0	0	0	0	1	0
$y_{.j}$	13	11	10	9	8	8	6	5	5	4	79	
$M^P_{.j}$	13	11	10	9	8	8	6	5	5	4	79	

and the caution index, C_j, for item j is given by

$$C_j = 1 - \frac{\sum_{i=1}^{N}(y_{ij} - P_{.j})(y_{i.} - P_{..})}{\sum_{i=1}^{N}(M^P_{ij} - P_{.j})(y_{i.} - P_{..})}$$

The second term of the caution index for item j is the ratio of two covariances: The numerator is the covariance of column vector j, (y_{ij}) and $(y_{i.})$, i=1,...,N and the denominator is the covariance of the vectors $(y_{i.})$ and (M^P_{ij}), i=1,2...,N. The value of the denominator is considered as a norm value to standardize the numerator.

It can be said that this ratio in the above caution index is equal to the ratio of the traditional discriminating index, r_j, total-item correlation to the standardized (or ideal in a sense illustrated in Table 2) discriminating index, r_j', for item j. That is

$$\frac{cov_j(y_{ij}, y_{i.})}{cov_j(M^P_{ij}, y_{i\cdot})} = \frac{\dfrac{cov_j(y_{ij}, y_{i.})}{\sigma_j(y_{ij})\,\sigma(y_{i.})}}{\dfrac{cov_j(M^P_{ij}, y_i)}{\sigma_j(M^P_{ij})\,\sigma\,(yi.)}} = \frac{r_j}{r_j'}$$

It is clear that $\Sigma(y_{ij} - P_{.j})^2 = \Sigma(M^P_{ij} - P_{.j})^2$ because the number of 1's in column j is invariant as can be seen in Tables 1 and 2, so the number of 1's in the column vector j, (M^P_{ij}) and (y_{ij}) are the same. Therefore, the two variances $\sigma^2_j(y_{ij})$ and $\sigma^2_j(M^P_{ij})$ are equal.

The Extended Caution Index in Conjunction With Response Theory Test and Group Response Curves: One Parameter Logistic Model.

According to the one parameter logistic model, the item response curve may be written

$$P_{b_j}(\theta) = \frac{1}{1+\exp[-D(\theta-b_j)]}, \quad j=1,2,\ldots,n \quad ,$$

where θ is the latent ability, b_j is the difficulty of item j and D is a constant which is set equal to -1.7 for convenience of comparison to the normal ogive model (see Lord & Novick, 1968, p.400). In the above equation, b_j is fixed and θ is a random variable.

Although in practice, the number of items, n, is a finite number, it is useful to consider b as a continuous variable. By holding θ_i fixed and treating b as a continuous variable, the dual function, $S_{\theta_i}(b)$, of the one parameter logistic function may be defined,

$$S_{\theta_i}(b) = \frac{1}{1+\exp[-D(\theta_i-b)]} \text{ , } i=1,2,\ldots,N \text{ .}$$

Of course, the expression

$$\frac{1}{1+\exp[-D(\theta_i-b_j)]}$$

may be considered to be a function of either θ or b. By choice of which variable is fixed, the function may be used to define either the item response curve, $P_{b_j}(\theta)$ or the person response curve $S_{\theta_i}(b)$ [see Lumsden, 1978, Weiss (1977)]. Hence, the variable described within the parenthesis of the function is considered as a random variable and the subscript variable is a fixed variable.

The curves for the pair of functions, $P_{b_j}(\theta)$ and $S_{\theta_i}(b)$ are symmetric about the vertical axis at $\theta = \theta_o$ (or equivalently $b = b_o$) provided $\theta_o = b_o$. As illustrated in Figure 1, however, the item response curve (IRC) and the person response curve (PRC), intersect at

$$(\theta_o + b_o)/2 \text{ if } \theta_o \neq b_o.$$

Insert Figure 1 about here

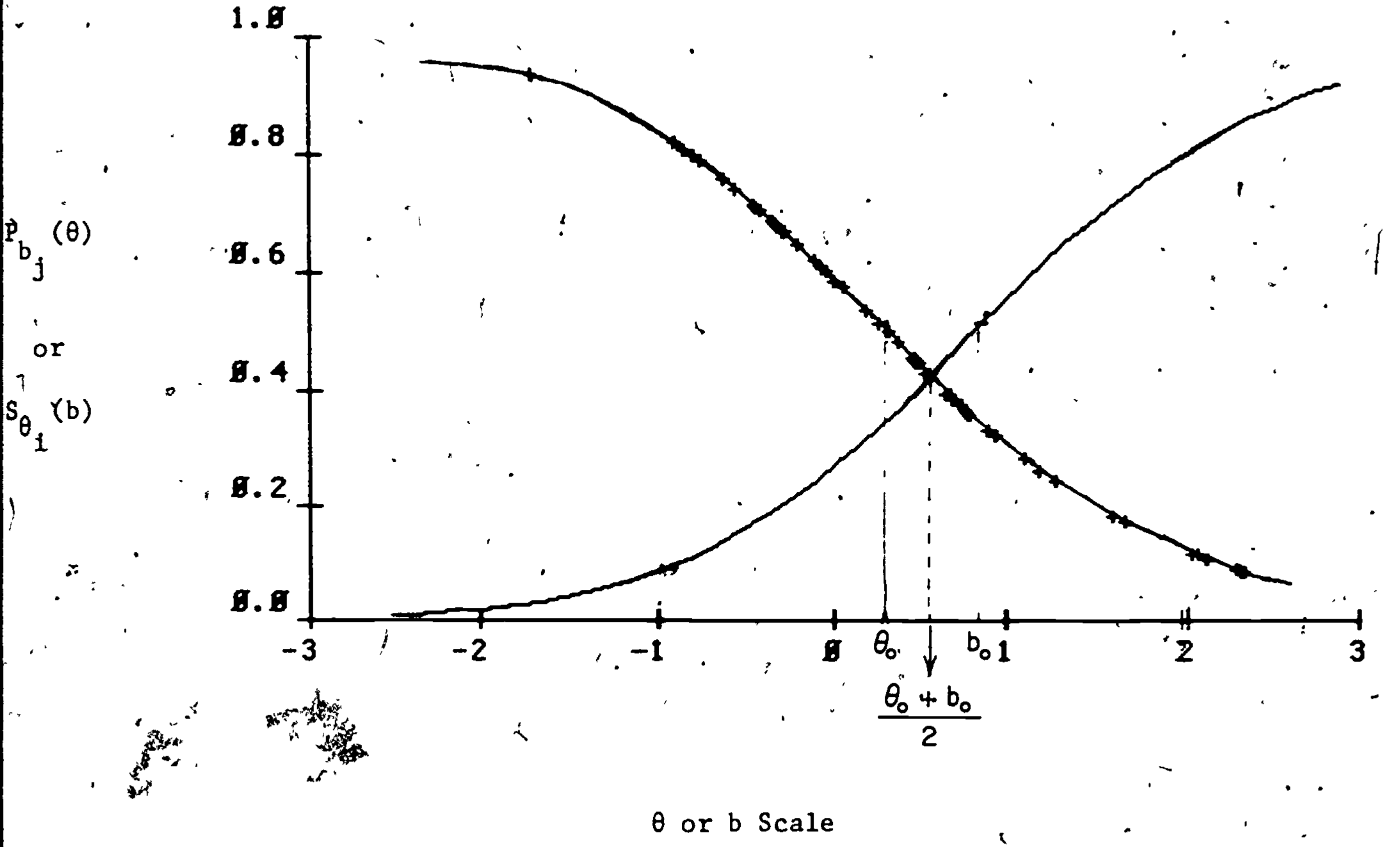

Figure 1. Person Response Curve (monotonically decreasing) and Item Response Curve (monotonically increasing) with the Mean Values of θ_o and b_o., One Parameter Logistic.

Addition and subtraction, an inner product of two functions in the same family (i.e. in one of the two families $\{P_{b_1}(\theta), P_{b_2}(\theta), \ldots P_{b_n}(\theta)\}$ or $\{S_{\theta_1}(b), \ldots S_{\theta_N}(b)\}$ in this paper), the norm of a function and the distance of any two functions in the same family will be defined below.

Definition. Addition and subtraction of two functions, $P_{b_1}(\theta)$ and $P_{b_2}(\theta)$, or $S_{\theta_1}(b)$, and $S_{\theta_2}(b)$ is defined as pairwise addition or subtraction of the two. That is,

$$(P_{b_1} \pm P_{b_2})(\theta) \equiv P_{b_1}(\theta) \pm P_{b_2}(\theta)$$

and

$$(S_{\theta_1} \pm S_{\theta_2})(b) \equiv S_{\theta_1}(b) \pm S_{\theta_2}(b)$$

Definition. An inner product (or the sum of the cross products) of the two functions is the sum of pairwise products $P_{b_1}(\theta_i)P_{b_2}(\theta_i)$ [or equivalently $S_{\theta_1}(b_j)S_{\theta_2}(b_j)$] or more generally, the integration of the product of the two functions with respect to θ (or b). Thus

$$[P_{b_1}(\theta), P_{b_2}(\theta)] = \sum_{i=1}^{N} P_{b_1}(\theta_i)P_{b_2}(\theta_i)$$

$$\text{or} \quad = \int P_{b_1}(\theta)\, P_{b_2}(\theta)d\theta$$

$$\text{and } [S_{\theta_1}(b), S_{\theta_2}(b)] = \sum_{j=1}^{n} S_{\theta_1}(b_j)S_{\theta_2}(b_j)$$

$$\text{or} \quad = \int S_{\theta_1}(b)S_{\theta_2}(b)db .$$

Definition. The squared norms of functions $P_b(\theta)$ and $S_\theta(b)$ are given by the inner product of themselves. Thus, we have

$$||P_b||^2 = [P_b(\theta), P_b(\theta)]$$

$$= \sum_{i=1}^{N} P_b{}^2(\theta_i) \text{ or } \int P_b{}^2(\theta)d\theta \; ,$$

and

$$||S_\theta||^2 = [S_\theta(b), S_\theta(b)]$$

$$= \sum_{j=1}^{n} S_\theta(b_j) \text{ or } \int S_\theta{}^2(b)db.$$

Definition. The squared distance of two functions $P_{b_1}(\theta)$ and $P_{b_2}(\theta)$ [or $S_{\theta_1}(b)$ and $S_{\theta_2}(b)$] is the inner product of their difference,

That is

$$||P_{b_1} - P_{b_2}||^2 =$$

$$= [P_{b_1}(\theta) - P_{b_2}(\theta), P_{b_1}(\theta) - P_{b_2}(\theta)]$$

$$= ||P_{b_1}||^2 + ||P_{b_2}||^2 - 2(P_{b_1}, P_{b_2})$$

and

$$||S_{\theta_1} - S_{\theta_2}||^2 =$$

$$= [S_{\theta_1}(b) - S_{\theta_1}(b), S_{\theta_1}(b) - S_{\theta_2}(b)]$$

$$= ||S_{\theta_1}||^2 + ||S_{\theta_2}||^2 - 2(S_{\theta_1}, S_{\theta_2})$$

By using the notation of integration,

$$||P_{b_1} - P_{b_2}||^2 = \int [P_{b_1}(\theta) - P_{b_2}(\theta)]^2 \, d\theta$$

or

$$||S_{\theta_1} - S_{\theta_2}||^2 = \int [S_{\theta_1}(b) - S_{\theta_2}(b)]^2 \, db.$$

With these definitions, we are ready to introduce the dual concept of Test Response Curve (Lord, 1980; Lord and Novick, 1968). This is the Group Response Curve as an average function of N different Person Response Curves. The Test Response Curve (TRC) is an average function of n IRC's defined as

$$T(\theta) = (1/n) \sum_{j=1}^{n} P_{b_j}(\theta).$$

Similarly, the Group Response Curve (GRC) is an average function of N PRC's, that is,

$$G(b) = (1/N) \sum_{i=1}^{N} S_{\theta_i}(b).$$

Illustrative PRC's and IRC's for 100 hypothetical persons were generated by randomly sampling 100 values of θ from a unit normal distribution. The resulting TRC for the simulated 100 item test is shown as the monotonically increasing function in Figure 2.

Insert Figure 2 about here

The curve that is a monotonically decreasing function is the PRC of $\theta = 0$, denoted by $S_0(b)$. The curve represented by "+"s is a Group Response

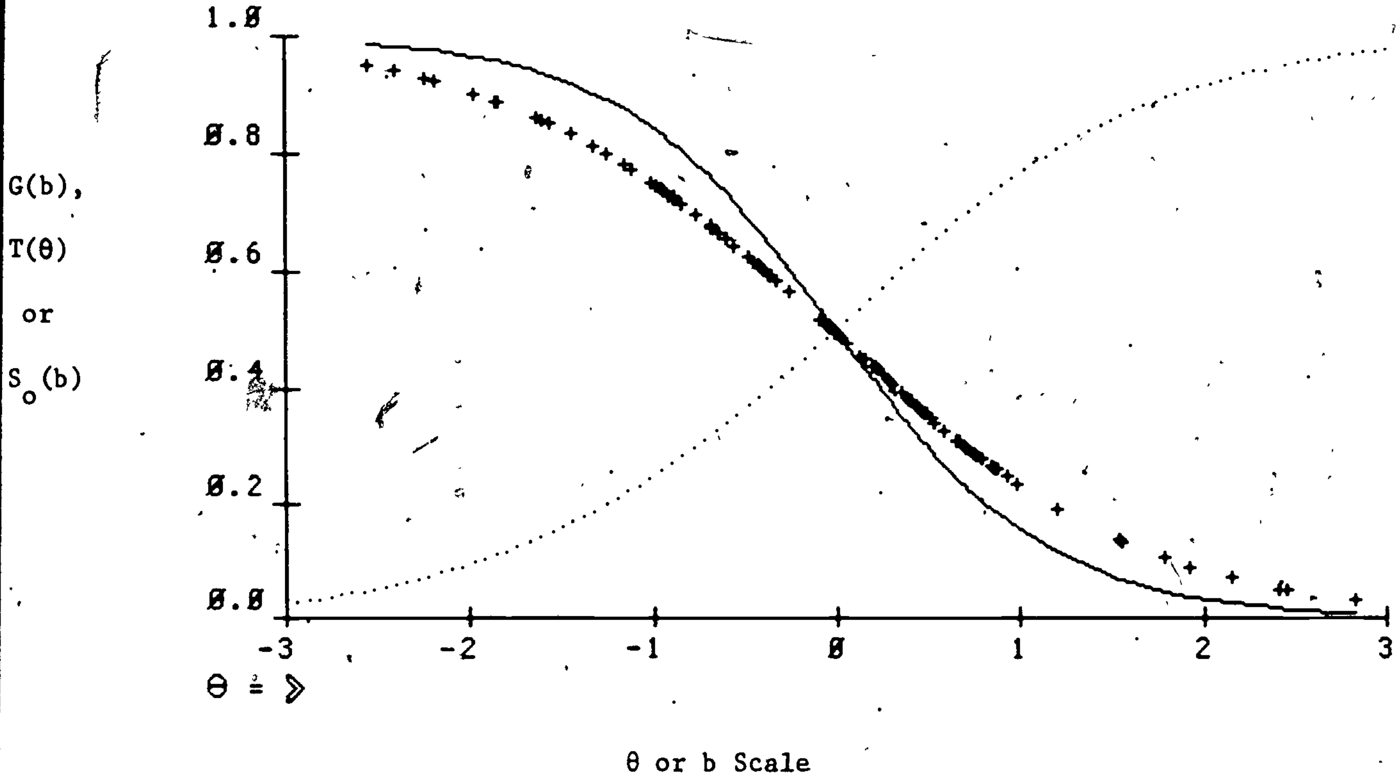

Figure 2. Test Response Curve (... line), Group Response Curve (+++ line) and Person Response Curve (solid line) of the One Parameter Logistic Model.

Curve which is obtained by taking the pointwise mean of 100 PRC's over the randomly generated 100 b values. That is,

$$G(b) = \frac{1}{100} \sum_{i=1}^{100} S_{\theta_i}(b)$$

As the number of b values approaches infinity, then G(b) in the figure will be a smooth curve, monotonically decreasing and moreover, if the number of θ values is also very large then G(b) will be a symmetric curve of T(θ) about the vertical line of θ = b = 0. With this figure, θ_i, i=1,2,...100 and b_j, j=1,2,...100 are randomly chosen from N(0,1) so their means are not exactly zero. It can be shown numerically that T(θ) and G(b) reach 1/2 at $\bar{\theta} = \frac{1}{100} \sum_i \theta_i$ and $\bar{b} = \frac{1}{100} \sum_j b_j$ respectively.

Let us denote the average of $T(\theta_i)$, i = 1,...,N by T,

$$T = (1/N) \sum_{i=1}^{N} T(\theta_i)$$

and the average of $G(b_j)$, j=1,...,n by G,

$$G = (1/n) \sum_{i=1}^{n} G(b_j).$$

Then T = G, because

$$\begin{aligned} T &= (1/N) \sum_{l=1}^{N} T(\theta_i) \\ &= (1/nN) \sum_j \sum_i \{1/1+\exp[-D(\theta_i - b_j)]\} \\ &= (1/n) \sum_{j=1}^{n} G(b_j) = G \end{aligned}$$

Definition of Various Extended Caution Indices

Sato's (1975) S-curve may be viewed as a discrete test response curve. The perfect S-curve divides 1's and 0's into two mutually exclusive areas with 1's under the curve and 0's above it. Note, however, that direct correspondence in this way involves a reordering of the subjects from low to high rather than from high to low as typically presented by Sato and as was shown in Table 2. represents the average probability of correctly answering items on the test when a person's ability is equal to θ. The analogy between the S-curve and a TRC may be seen by considering an alternative N by n score matrix with real numbers based on IRT rather than binary scores. More specificaly, let

$$PM_{ij} = P_{b_j}(\hat{\theta}_i)$$

where $\hat{\theta}_i$ is an estimated ability parameter, θ, for person i and $\hat{b}_j$ estimated item parameter for item j under the condition that

$$\sum_{j=1}^{n} P_{b_j}(\hat{\theta}_i) = \sum_{j=1}^{n}$$

Since $$P_{b_j}(\theta_i) = S_{\theta_i}(b_j)$$

for fixed i and j, the cells of the probability matrix (PM_{ij}) are also equal to $S_{\hat{\theta}_i}(\hat{b}_j)$. If the rows and columns of this matrix are arranged in the manner of the S-P table and columnwise sums of the cell entries are obtained, the result is N times $G(b_j)$, which corresponds to the P-curve. Similarly, n times $T(\theta_i)$ corresponding to the S-curve may be obtained by summing the cell entries for each row.

Selected rows and columns of a probability matrix (PM_{ij}) are illustrated in Table 3 for a 32 item test involving the subtraction of signed numbers that was administered to a sample of 127 students (Tatsuoka & Tatsuoka, 1981). Also shown in Table 3 are the values of the estimated item and ability parameters and the test and group response curves evaluated at those estimated parameter values (i.e., $T(\theta_i)$ and $G(b_j)$ respectively).

Table 3

The 127 x 32 Probability Matrix (PM_{ij}) for

Signed-Number Subtraction Problems

item j / $S\#_i$	1	2	- -	15	16	- -	31	32	$T(\hat{\theta}_i)$	$\hat{\theta}_i$
1	.000	.001	· · ·	.040	.002	· · ·	.017	.082	.026	-1.114
2	.000	.002	· · ·	.061	.004	· · ·	.035	.129	.038	-0.916
3	.000	.002	· · ·	.061	.004	· · ·	.035	.129	.038	-0.722
	⋮	⋮		⋮	⋮		⋮	⋮		
60	.549	.635	· · ·	.783	.871	· · ·	.969	.969	.809	.700
61	.567	.647	· · ·	.789	.878	· · ·	.970	.970	.816	.710
62	.568	.648	· · ·	.789	.878	· · ·	.970	.970	.817	.714
	⋮	⋮		⋮	⋮		⋮	⋮		
88	.860	.854	· · ·	.832	.962	· · ·	.994	.998	.968	1.222
	⋮	⋮		⋮	⋮		⋮	⋮		
127	1.000	1.000	· · ·	1.000	1.000	· · ·	1.000	1.000	1.000	+ ∞
$G(\hat{b}_j)$	.527	.570		.691	.708		.837	.843		
$\hat{b}_j$	-.467	-.467		-.044	.021		.289	.378		

Insert Table 3 about here

Before introducing the extended caution index, it is useful to compare the S and P curves for the data from which the estimates in Table 3 were obtained with their counterparts, i.e., n times $T(\theta_i)$ and N times $G(b_j)$. The two comparisons S with $nT(\theta_i)$ and P with $NG(b_j)$ are provided in Figures 3 and 4 respectively. The tic marks on the horizontal axis in Figure 3 indicate the location of the θ's for the 127 students in the study. The tic marks in Figure 4 show the values of b_j for the 32 items. The close correspondence between the two pairs of curves is apparent. The number of items and the limited range of values that $\hat{b}_j$ assumes for these data obviously limits the evaluation of the correspondence between the curve's in Figure 4, however.

Insert Figures 3 & 4 about here

Given the parallels between the S-P curves and the GRC and TRC, the extension of the caution index for use with the latter curves is relatively straightforward. There are, however, several natural ways in which the extension can be made. Possibly the most obvious extension is to simply replace the term $(M^s_{ij} - P_{i.})$ in the denominator of equation (1) by its counterpart from the PM_{ij} matrix, i.e.,

$$[PM_{ij} - T(\hat{\theta}_i)] = [S_{\hat{b}_j}(\hat{\theta}_i) - T(\hat{\theta}_i)].$$

With the above substitution, our first extended caution index, $C1_i$, is defined

$$C1_i = 1 - \frac{\sum_j (y_{ij} - P_{i.})(y_{.j} - P_{..})}{\sum_j [S_{\hat{\theta}_i}(\hat{b}_j) - T(\hat{\theta}_i)](y_{.j} - P_{..})}$$

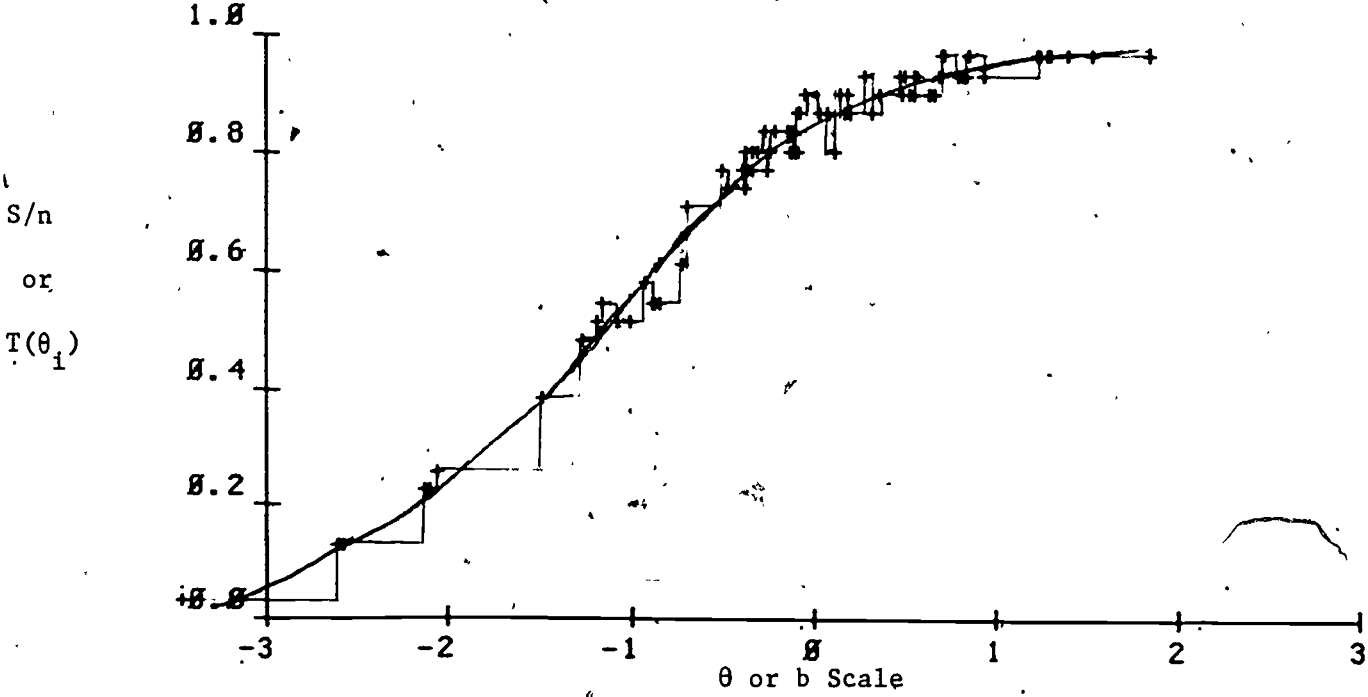

Figure 3. Comparison of S-curve (Converted to the Proportion Correct to Subject) With the Test Response Curve for the Data in Table 3.

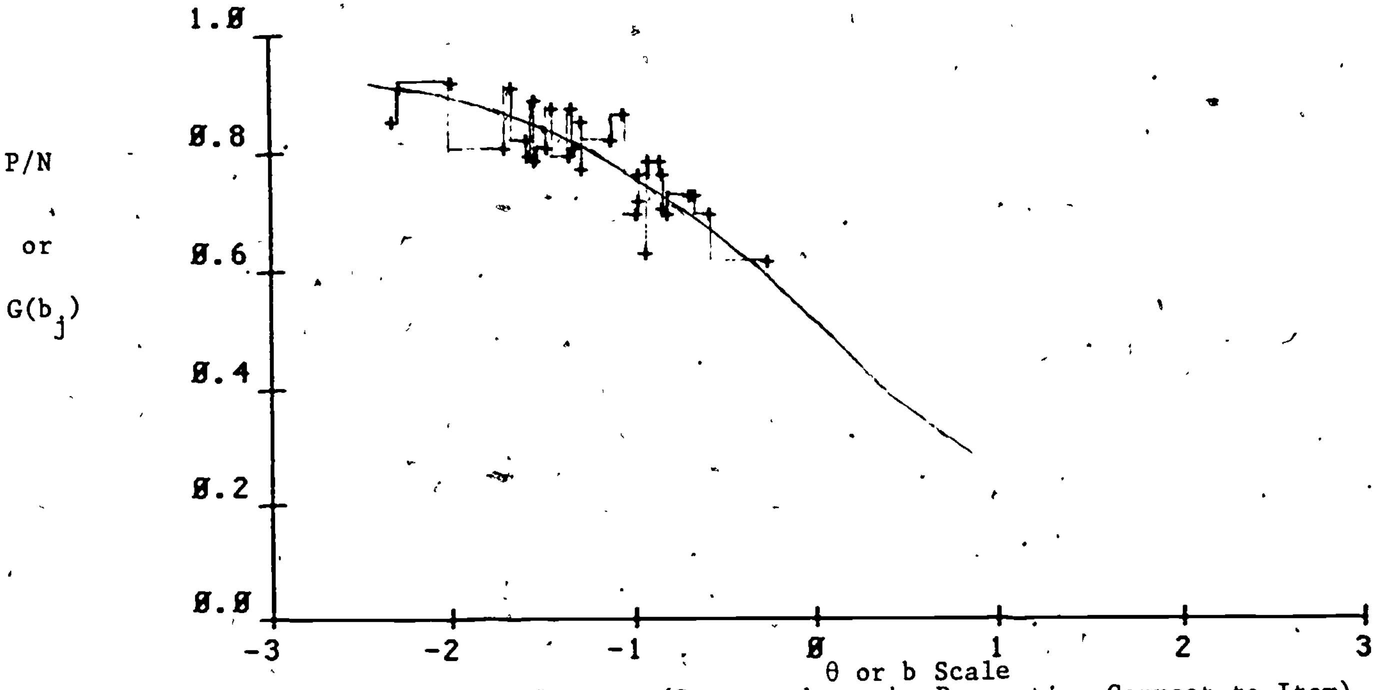

Figure 4. Comparison of P-curve (Converted to the Proportion Correct to Item) With tne Group Response Curve for the Data in Table 3.

The numerator divided by n, i.e., the covariance of (y_{ij}) and $(y_{.j})$, can be expanded to the sum of

$$(1/n) \sum_{j=1}^{n} y_{ij}y_{.j} \text{ and } -P_{..}P_{i.}$$

The value of the second term does not depend on a person's response to each item but depends on his/her total score. As long as the total score is fixed, the anomaly of response patterns will not be detected by this value. This value varies between persons, so if two persons have the same achievement level θ_i, then the judgment regarding the extent to which each response pattern deviates from the norm depends only on the first term of the numerator. Since the denominator is a normalized constant for a fixed value, $\hat{\theta}_i$, it is unlikely that a particular aberrant response pattern produced by an individual whose achievement level is θ_i will affect the denominator.

Thus, it is natural to expect that if both the quantities are replaced by the inner products of the two row vectors (y_{ij}) and $(y_{.j})$ for j=1, 2,..., n, the values of $C1_i$ will be affected by the degree of anomaly of individual response patterns. Moreover, calculation of inner products is easier than that of covariances. Let us define four other natural extentions of the Caution index as follows.

Definition. Four alternative definitions of the extended caution index for person i are:

$$C2_i = 1 - \frac{\sum_{j=1}^{n} y_{ij}y_{.j}}{\sum_{j=1}^{n} S_{\hat{\theta}_i}(\hat{b}_j)y_{.j}} ,$$

$$C3_i = 1 - \frac{\sum_{j=1}^{n} y_{ij}S_{\hat{\theta}_i}(\hat{b}_j)}{\sum_{j=1}^{n} G(\hat{b}_j)S_{\hat{\theta}_i}(b_j)} ,$$

$$C4_i = 1 - \frac{\sum_{j=1}^{n} y_{ij} G(\hat{b}_j)}{\sum_{j=1}^{n} S_{\hat{\theta}_i}(\hat{b}_j) G(b_j)} ,$$

and

$$C5_i = 1 - \frac{\sum_{j=1}^{n} y_{ij} S_{\hat{\theta}_i}(\hat{b}_j)}{\sum_{j=1}^{n} y_{ij} G(\hat{b}_j)} .$$

The denominators of the four indices are considered as normalizing constants but the characteristics of the numerators will be divided into two categories. The indices in the first category, $C2_i$ and $C4_i$ give measures that are more group dependent, because they are the sums of cross products of the corresponding elements of the observed vector $(y_{ij})'$ and the row-sum total vector $(y_{.j})$, and Group Response Curve $G(b_j)$ respectively. They measure the relationship of an observed response pattern for a person i to a normed variable derived from the group the person i belongs to. Thus these indices have a similiar function to the Norm Conformity Index, NCI, defined in Tatsuoka & Tatsuoka (1980). The remaining indices, $C3_i$ and $C5_i$, are more individually oriented. That means the quantities obtained from $C3_i$ and $C5_i$ reflect the extent a person i's response pattern (y_{ij}) relates to a theoretically derived PRC at the fixed level of θ_i. Thus, it can be said that the indices $C3_i$ and $C5_i$ are similar to the Individual Consistency Index (Tatsuoka & Tatsuoka, 1980).

These extended caution indices for person i will be easily altered to those for item j.

$$C2_j = 1 - \frac{\sum_{i=1}^{N} y_{ij}\, y_{i.}}{\sum_{i=1}^{N} P_{B_j}(\hat{\theta}_i) y_{i.}} ,$$

$$C3_j = 1 - \frac{\sum_{i=1}^{N} y_{ij} \; P_{b_j}(\hat{\theta}_i)}{\sum_{i=1}^{N} P_{\beta_j}(\hat{\theta}_i)T(\hat{\theta}_i)} ,$$

$$C4_j = 1 - \frac{\sum_{i=1}^{N} y_{ij} \; T(\hat{\theta}_i)}{\sum_{i=1}^{N} P_{\beta_j}(\hat{\theta}_i)T(\hat{\theta}_i)} ,$$

and

$$C5_j = 1 - \frac{\sum_{i=1}^{N} y_{ij} \; P_{\beta_j}(\hat{\theta}_i)}{\sum_{i=1}^{N} y_{ij} \; T(\hat{\theta}_i)} .$$

Similarly, the indices $C3_j$ and $C5_j$ are potentially useful for detecting anomalous response patterns in comparison with item j's IRC while $C2_j$ and $C4_j$ are potentially useful indices for purposes of identifying items of which patterns deviate from that of test, TRC.

The Case of Two and Three Parameter Logistic Models

Problems in Person Response Curves and Group Response Curves

Person Response Curves for the one parameter logistic model are represented by smooth monotonically decreasing functions defined over the difficulties of the infinitely many items. But PRC for the two parameter logistic model is no longer a smooth, monotonically decreasing curve. Figure 5 provides the graph of Person Response Curve for the ability levels of $\theta = 0$ as well as Test Response Curve of the two parameter logistic model where ability measures θ_i, i=1,2,...,100, were randomly sampled from a normal (0,1) distribution, the difficulties b_j, j=1,2,...,100 were also randomly sampled from a normal (0,1) distribution and the item discrimination indices, a_j j=1,...,100, were drawn from the uniform distribution of the interval (0.8, 1). Test Response Curve, Person Response Curves are given by

$$T(\theta) = (1/n)\sum_{j=1}^{n} P_{b_j}(\theta)$$

and

$$S_{\theta_o}(b) = \frac{1}{1+\exp[-Da(\theta_o-b)]}$$

for a fixed θ_o and variable b

Insert Figure 5 about here

The dotted line (+++) in the figure is the Group Response Curve of a hundred subjects. Although each PRC is locally oscillated, especially around the origin, the GRC (the mean curve of these PRCs) becomes fairly smooth and almost monotonically decreasing. Since b_j, j=1,...100 are randomly selected from N(0,1), a larger oscillation of PRC around the mean 0 is expected. But GRC is expected to be smoother as the number of students and items increase to a larger number.

Insert Figure 6 about here

Figure 6 is the graph of TRC, GRC, PRC of $\theta = 0$ for the three parameter logistic model. The parameters θ_i, b_j and a_j were generated by the same method as that of the two parameter model then fifty C-values of 0.15, and 50 of 0.20 were randomly assigned to 100 pairs of a_j and b_j to make the three parameter logistic model.

It seems that the smoothness of the curve GRC for three parameter logistic model is about the same, differing only as expected in terms of the lower asymptote. A larger number of subjects will be needed for the three parameter case in order to obtain smoother GRC.

The definition of the extended caution indices may be applied more generally to the two and three parameter logistic models in essentially

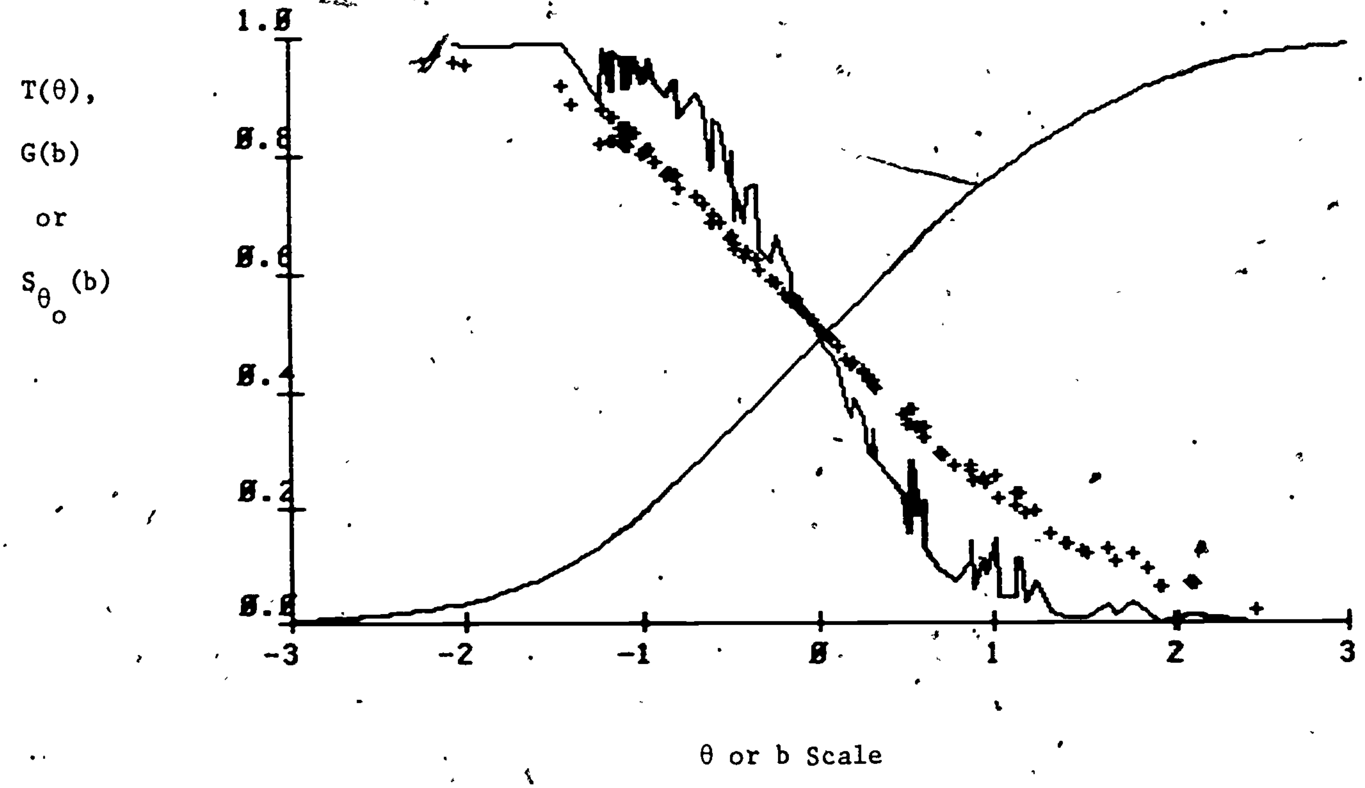

Figure 5. Test Response Curve (solid line), Group Response Curve (+ + line) and Person Response Curve (jagged line) of Two Parameter Logistic.

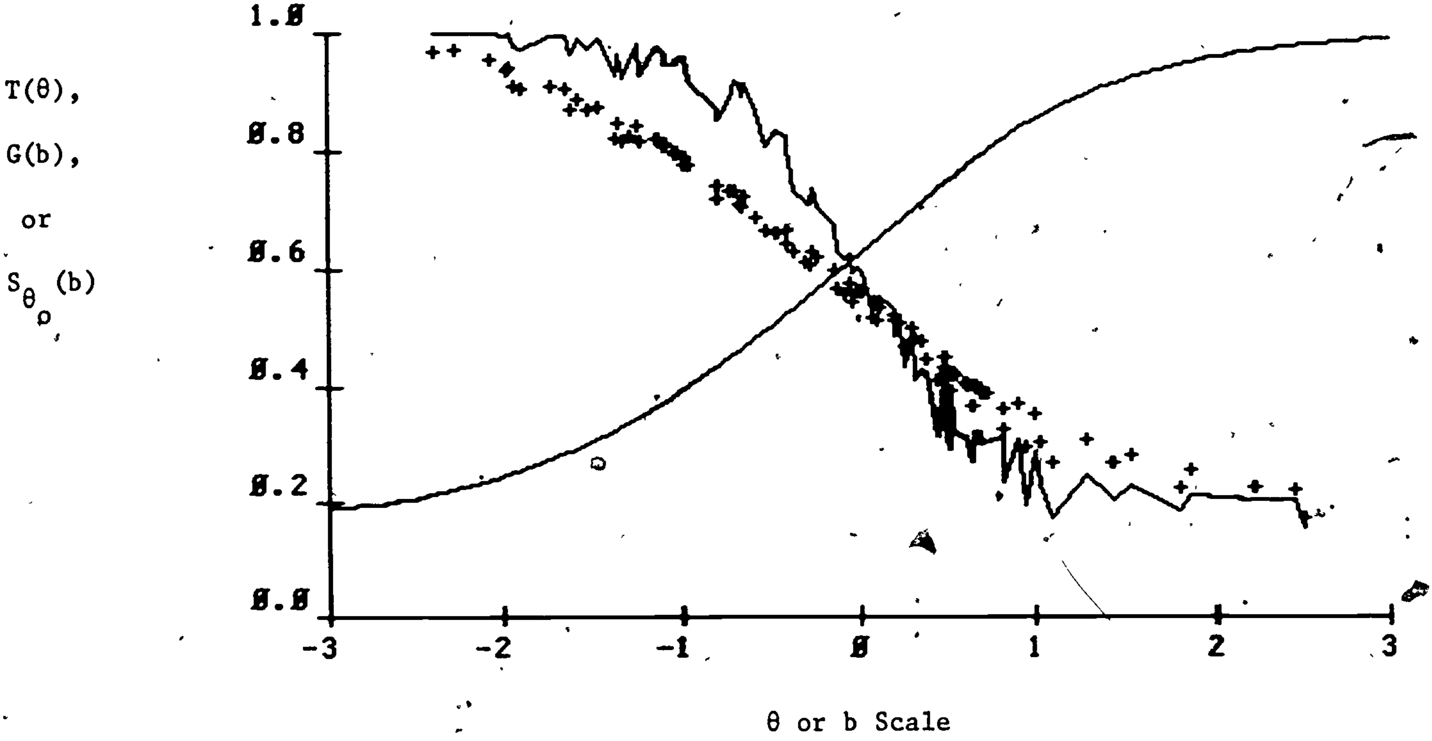

Figure 6. Test Response Curve (solid line), Group Response Curve (++ line) and Person Response Curve (jagged line) of Three Parameter Logistic.

the same manner as it was develoed for the one parameter model.

Note that the arrangement of rows and columns according to the orders of the proportion, correcčts (p values) for n items and the total scores for N subjects is essential to determine S-P curves, and the values of M^p and M^s, i=1,2,...,N, j=1,...,n. With our extended caution indices, the arrangements of rows and columns in monotonic order of the probability are no longer necessary.

Application of New Indices for the Detection of Anomalous Responses.

There is evidence that student errors on certain types of arithmetic problems are frequently quite systematic (Brown and Burton, 1978; Birenbaum and Tatsuoka, 1980 Davis, McKnight, 1980). That is, students seem to consistently apply erroneous algorithms in attempting to answer a problem of a particular form. Sometimes erroneous or incomplete rules result in the right answer. For example, a student who consistently treats a multiplication sign as if it were an addition sign would get the right answer to the problem 2 x 2 = 4, but would get it for the wrong reason. A score of zero for using the wrong operation would be a better reflection of the student's ability to multiply than a score of one for answering "4" to the item.

Birenbaum and Tatsuoka (1980) have demonstrated that the customary zero-one scoring of incorrect and corrent answers can give the appearance of higher dimensionality and cause difficulty in attempting to apply IRT when students consistently apply erroneous rules to the addition and subtraction of signed numbers. The difficulties result from the fact that several erroneous rules frequently yield the right answer for some problems. Right answers for the wrong reasons not only cause problems in applying IRT, but more importantly they can result in misleading scores and make it difficult to diagnose what a student is doing wrong.

By painstaking work Tatsuoka and her colleagues (Birenbaum and Tatsuoka, 1980; Birenbaum, 1981) were able to identify several erroneous rules that were consistently applied by certain students. Birenbaum and Tatsuoka (1980) reanalyzed their data after converting ones to zeroes for items that students got right for the wrong reasons. That is, an item score was changed from one to zero if (1), a student was identified as consistently applying an erroneous rule and (2) application of that erroneous rule would lead to the correct answer for the particular item in question. Analysis of the resulting modified data indicated that the data were more nearly unidimensional and there was good evidence that IRT was more applicable to the modified data than to the original data.

Anomalous response patterns can sometimes be found by conducting an intuitive error analysis or by clinical interviews. Both approaches require enormous effort. Brown and Burton (1978) and Tatsuoka et al. (1980) have developed cumputerized approaches to error analysis. But these methods are expensive and were based on extensive work with highly specific item content.

Tatsuoka and Tatsuoka (1981) demonstrated an index, called the individualized consistency index (ICI) which was shown to be useful in detecting a variety of erroneous rules of operation of signed-number addition and subtraction problems. Using the ICI to detect examinees who are apt to have a misconception saves considerable effort because only examinees so identified have their item responses routed to the detailed error-diagnostic system. Application of the ICI is limited, however, because it requires repeated measures, i.e., several items based on an identical item form, within the test. Such repetition is not common on most tests.

As will be seen below, the index similar to ICI, $C3_i$, not only avoids the repeated measure limitation but is apparently more effective for purposes of detecting anomalous response patterns resulting from the consistent application of an erroneous rule. Tatsuoka & Tatsuoka (1981) showed a list of erroneous rules of operation ("bugs") detected by ICI. The 32 response patterns resulting from these bugs are classified in Group A. The rest of the 103 response patterns are classified into two groups according to the error-diagnostic system, SIGNBUG. Group B consists of 7 responses which are probably using one or two erroneous rules inconsistently; Group C, responding adequately using the right rule of operation and/or no indication of systematic errors. The errors observed in Group C are apparently just random errors. The estimated item and person ability parameters needed to compute the extended caution indices were obtained by the computer program GETAB (Robert Baillie, 1979), using Birenbaum & Tatsuoka's modified dataset.

Distributions of the indices $C2_i$ and $C3_i$ are displayed in Figures 6 and 7 respectively. Only members of groups A and B (persons who consistently used an erroneous rule) and of group C (persons who made a substantial number of errors but whose errors were not the result of consistent use of an erroneous rule) are included in the distributions shown in Figures 6 and 7. In both figures, persons in group A and B are depicted by shaded boxes and those in group C by unshaded boxes.

Insert Figures 7 and 8 about here

As can be seen in Figure 6, $C2_i$ does not provide any basis for distinguishing persons who are consistently using an erroneous rule from those who aren't. The two groups are distinguished almost perfectly, however, by the magnitude of $C3_i$ (see Figure 7). Indeed, there is almost no overlap between the two groups. All 39 members of Groups A and B have values of $C3_i$ of .05 or higher whereas only two of the 88 members of group C have positive values of $C3_i$ and the rest of the members of group C have values of $C3_i$ as large as .05. Thus, $C3_i$ may be used to identify with a high degree of accuracy those persons who consistently use an erroneous rule.

As might be expected from a comparison of the coefficients, $C4_i$ works in a fashion quite similar to $C2_i$, and $C5_i$ works much like $C3_i$ in terms of the abiliy of these indices to distinguish members of groups A, B and C. It is clear that $C2_i$ and $C4_i$ are not useful for detecting anomalous response patterns resulting from consistent application of an erroneous rule. These indices may be useful for other tasks for which NCI or Van de Flier's index (Harnisch & Linn, 1981) have been found to be useful. The third and fifth indices ($C3_i$ and $C5_i$) however, are quite effective for purposes of detecting persons who make consistent errors.

Insert Table 4 about here

Table 4 shows a summary of t-statistics comparing the means on the four generalized caution indices and ICI in the two groups: A and B combined versus C by itself. The t-value for index 2 is not significant

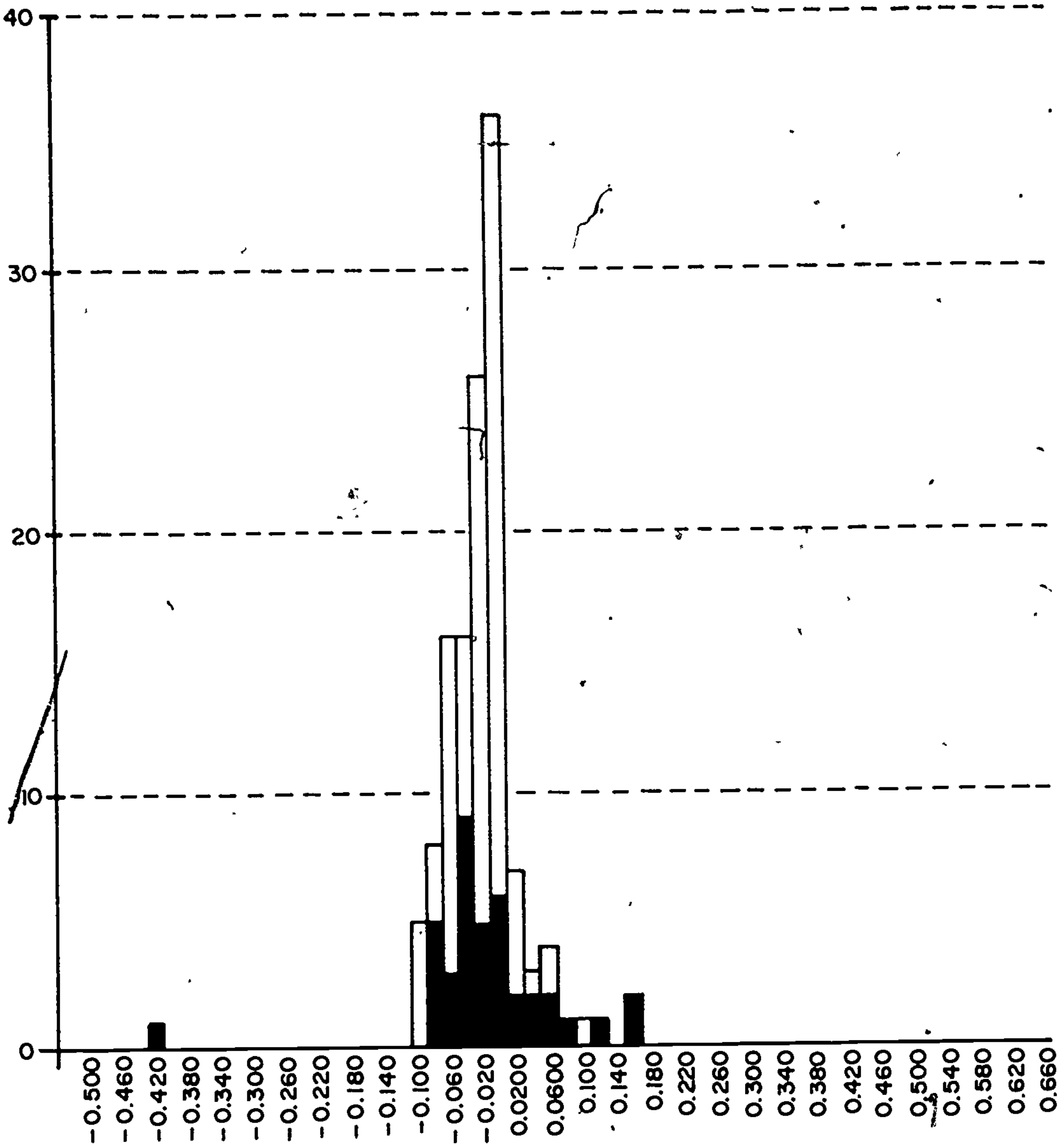

Figure 7. Histogram of Index $C2_i$: The shaded Area Represents the Members of Group A and B, -- Using Some Erroneous Rules, N=127

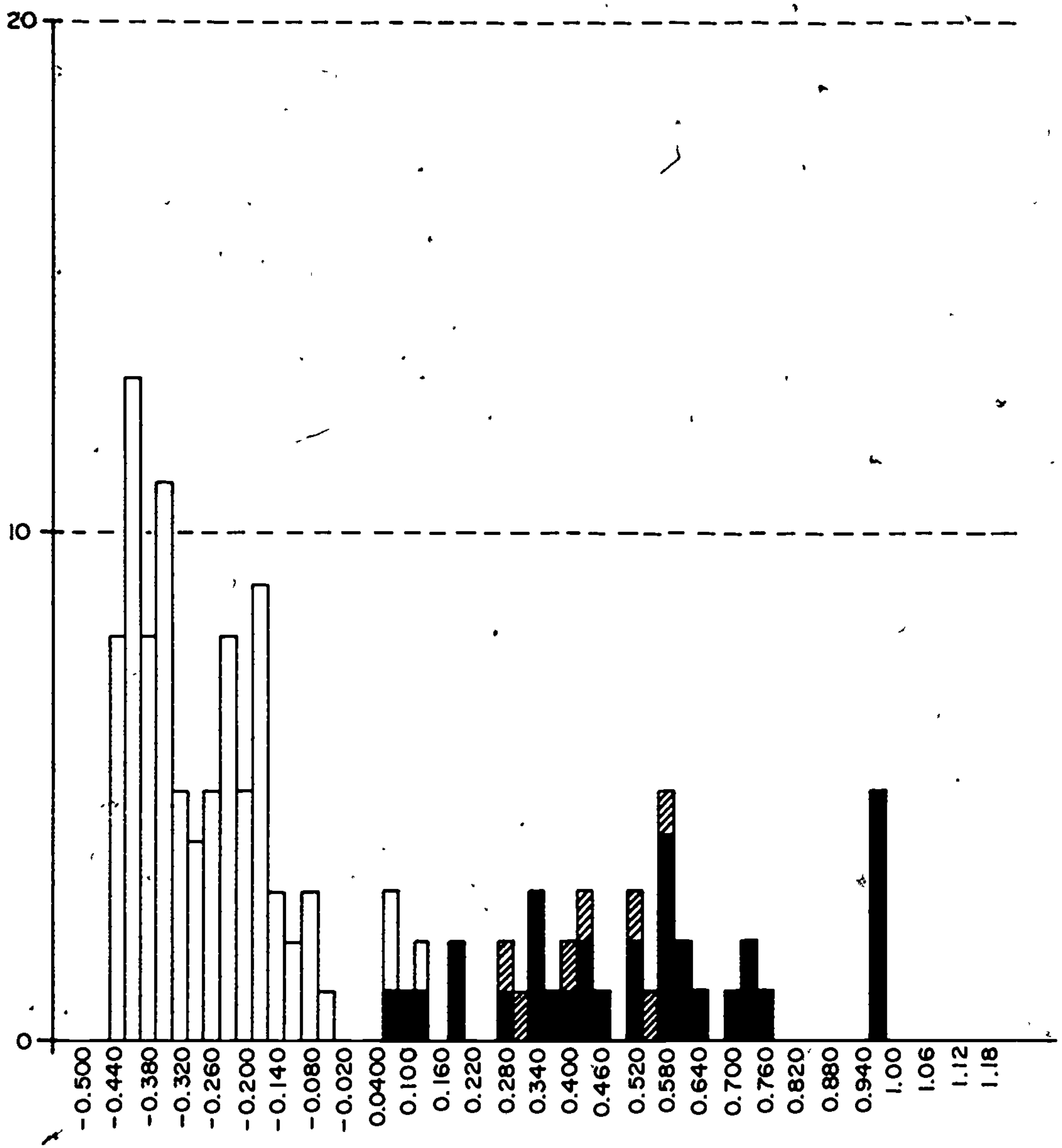

Figure 8. Histogram of Index $C3_i$: The Black Area Represents the Members of Group A and the Shaded for Group B -- Using Some Erroneous Rules, The White Area is for Those Using the Right Rule. N=127

Table 4

A Summary of t-statistics Comparing the Means on the Four Generalized Caution Indices and ICI in the Two Groups

Indices		Group A & B	Group C	t-value	P
	N	39	88		
Index 2	Mean	-.0170	-.0065		
	S.D.	.0929	.0306	.689	.4980
Index 3	Mean	.5310	-.2688		
	S.D.	.2444	.1300	-19.293	< .00005
Index 4	Mean	.0650	-.0045		
	S.D.	.1237	.0293	-3.466	< .0015
Index 5	Mean	.5091	-.2643		
	S.D.	.2615	.1350	-17.467	< .00005
ICI	Mean	.9223	.8144		
	S.D.	.0645	.1058	-7.121	< .00005

but all others are significant. Index 1 is excluded in the analysis because the denominator of this index becomes infinity when all items are correctly answered by all examinees.

Discussion

As was shown above, the caution index which Sato developed based solely on a comparison of observed item responses to group responses can be readily extended to theory based estimates of person and group response probabilities. The caution index is a linear transformation of the covariance of a person's response pattern with one or another theoretical curves computed using item-response theory. Alternatively, the extended caution indices may be viewed as linear transformations of the distance bewteen a person's response pattern and a theoretical curve (either the person response curve, as in the case of $C3_i$ and $C5_i$ or the group response curve, as in the case of $C4_i$).

The application of the extended caution indices that were introduced in this paper provided strong evidence that the indices that depend on the distance between a person's response pattern and their theoretical person response curve (i.e., $C3_i$ and $C5_i$) are quite effective for purposes of identifying persons who consistently use an erroneous rule in answering signed-number arithmetic problems. This is a potentially important result that deserves further investigation with other data sets involving different types of achievement test data. If additional research yields similar results, these indices may have considerable instructional utility because instruction can be made much more specific once it is determined that a student is consistently making an error as the result of a particular misconception.

References

Birenbaum, M. Error Analysis -- it does make a difference. Doctoral Dissertation, University of Illinois at Urbana-Champaign, 1981.

Birenbaum, M., & Tatsuoka, K. K. The use of information from wrong responses in measuring students' achievement (Research Report 80-1). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, 1980.

Brown, J. S., & Burton, R. R. Diagnostic models for procedural bugs in basic mathematics skills. Cognitive Science, 1978, 2, 155-192.

Davis, R. B., & McKnight, C. The influence of semantic content on algorithmic behavior. The Journal of Mathematical Behavior, 1980, 3, 39-87.

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: questionable test data and dissimilar curriculum practices. The Journal of Educational Measurement, 1981, in press.

Levine, M. V., & Rubin, D. B. Measuring the appropriateness of multiple-choice test scores. Journal of Educational Statistics, 1979, 4, 269-290.

Lord, F. M. Application of item response theory to practical testing problems. Hillsdale, N.J.: Erlbaum, 1980.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading: Addison-Wesley, 1968.

Lumsden, J. Tests are perfectly reliable. British Journal of Mathe- and Statistical Psychology, 1978, 31, 19-26.

Sato, T. The construction and interpretation of S-P tables. Tokyo: Meiji Tosho, 1975 (in Japanese).

Tatsuoka, K. K., & Tatsuoka, M. M. Detection of aberrant response patterns and their effect on dimensionality (Research Report 80-4). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, 1980.

Tatsuoka, K. K., & Tatsuoka, M. M. Spotting erroneous rules of operation by the Individual Consistency Index (Research Report 81-4). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, 1981.

Tatsuoka, K. K., Birenbaum, M., Tatsuoka, M. M., & Baillie, R. A psychometric approach to error analysis on response patterns (Research Report 80-3). Urbana, Ill.: University of Illinois, Computer-based Education Research Laboratory, 1980.

Trabin, T. E., & Weiss, D. J. The person response curve: fit of individuals to item characteristic curve models (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1979.

Wright, B. D., & Stone, M. H. Best test design, Rasch Measurement. Chicago: The University of Chicago, Mesa Press, 1979.